General Description

- This problem set uses data from the National Longitudinal Survey of Youth, 1997 Cohort. The data are described in further detail on a separate handout.

- The PIT problem set counts for 10% of your total grade for the course. Part I is worth 40% of the total PIT grade. Part II is worth 60% of the total PIT grade.

- In this problem set, you will:
  
  (1) get experience using a large data set;

  (2) respond to a specific set of questions regarding correlates of the health of youth (measured by their body mass index) of youth;

  (3) develop your own research question;

  (4) select key variables and appropriate methods to answer your questions;

  (5) deal with missing data;

  (6) apply the tools and concepts that you’ve learned in this course to gain insight into these questions;

  (7) identify the shortcomings of the data you are working with to answer the research question(s), acknowledging the strengths and limitations of your analysis.

  (8) convey your analysis in a concise, well-organized, and convincing manner.

Due Dates and Feedback Plan

- Part I is due in my GPPI mailbox by Monday, December 2, at 7 p.m.

- I will look at these drafts, make comments, assign a grade (see below) and return them to you by Thursday, December 5.

- Part II is due in my GPPI mailbox by Friday, December 13, at 5 p.m.
Your Tasks

PART I: WHAT FACTORS ARE ASSOCIATED WITH THE HEALTH OF YOUTH?

To answer this question, we’ll measure “health” using the body mass index (BMI). BMI is calculated as weight in kilograms divided by the square of the height in meters. While it is not a direct measure of health, overweight and obesity (measured by BMI) are known risk factor for, and associated with, a number of health problems. For further information, see, for example, http://www.niddk.nih.gov/health/nutrit/pubs/statobes.htm.

Write a memo to me, addressing the following:

(1) Provide basic descriptive statistics about the youth in the NLSY97: what percentage are female and male? what is the distribution of ages? what is their distribution across regions of the country? what is their race/ethnicity distribution? Select two other variables to include in this general descriptive section about the youth (not variables that you examine in the second part of Part I). Provide a written description in one to two paragraphs, using supporting tables, charts and/or graphs (note: clearly label these).

(2) Use appropriate methods to answer the following questions (you should not use the same method to answer more than one of the questions).

(a) Does youth BMI differ across the four regions of the country?

(b) Do male and female youth have the same level of BMI?

(c) Is family income associated with youth’s BMI?

(d) Is youth’s cigarette smoking in the past 30 days related to his or her BMI?

PART II: YOUR QUESTION HERE!

Using the variables in the data set, develop your own research question, similar to the types asked in part I (i.e., a very general question, and then investigate specific relationships of interest that relate to that question. A common variable does not need to be used throughout all four parts).

Your analysis must include descriptive statistics for the variables / issues you address, and you must appropriately use at least four of the following topics / tools covered in this course.

-- t-test for a single sample mean (Z-test for sample proportion)
-- t-test for difference of sample means / proportions
-- ANOVA
-- Chi-square tests and corresponding measures of association
-- Correlation (and possibly bivariate regression)
You are not expected to exhaust all possible investigations of interest: this is impossible given the time and resources available to you, and this is not meant to be a mini-practicum. The key is to specify a question, select appropriate data and methods to analyze that question, and convey your analysis in a clear and appropriate way.

Guidelines on writing memos

- I will provide an additional handout on appropriate memo style and format.

- Your analysis for Part I and Part II should be in memo form with:
  
  -- a brief introductory /overview paragraph that describes the overall research question and has a *brief* review of your findings.
  
  -- A paragraph or two (and any supporting tables) that convey descriptive statistics for the variables you use in the analysis (note: this would be an appropriate place in Part I to talk about the whole NLSY97 sample; you don’t need to do this entire-sample description for Part II).
  
  -- A separate paragraph (or two) for each of the questions/analyses (referring to supporting tables or figures). These should include a brief description of each question, including your informal expectations, and the formal null and alternative hypotheses you are testing; a brief description of your analytical approach (including a specific statement about the kind of statistical test you used) and findings (including discussions of both statistical significance (with appropriate statistics) and substantive significance).
  
  -- Discussion of the group(s) to which your analyses apply. i.e., to what population(s) do your findings generalize? This information may either be included in the paragraph for each question you address, or in a separate paragraph.
  
  -- A brief concluding paragraph.
  
  -- Tables, charts, graphs, figures should be appropriately labeled and formatted, and referred to at specific points in the body of the memo.

- There is no strict page or word limit. However, as a guideline, I think that you can complete your analysis in the equivalent of about 3 single-spaced pages of text. Your memo may be longer or shorter, depending on your use of bullet points, spaces, number of questions you address, etc.

- Do not turn in SAS output with your memo. It will be your responsibility to convey your findings and analyses through your memo, and in supporting tables or graphs that you
create (do not cut and paste SAS output: format the information in tables, graphs, and figures as you would for a “real” report).

- A note on audience for your memo: Your memo should be accessible to TWO types of readers:

  (1) readers who are not familiar with the language of statistics (e.g., an intelligent reader of a newspaper who has not had a class in statistics); and

  (2) readers who will only believe statements you make if you show that you arrived at your answer using statistics (e.g., your instructor for this course)! How do you balance these two audiences?

  -- For audience (1), it is your responsibility to appropriately interpret and translate your statistical analyses into language that is accessible to those who are not trained in statistics. This means being able to convey the gist of the “null hypothesis,” “statistical significance,” “research significance,” “generalizability,” etc. without actually using the terms (if you do use them, you must clearly explain them – not in stats jargon).

  -- For audience (2), you will want to include information about the type of test you conducted, the test statistics you obtained, and p-values (perhaps in parentheses, or as supporting information to the statements you make for audience (1)).

**Assessment Criteria**

Memos for Parts I and II will be assessed using the following criteria:

-- Addressed all parts of the assignment.
-- Selection of relevant variables (recoding if necessary) to answer each question.
-- Appropriate and effective application of statistical tools.
--- Appropriate and accurate conclusions drawn from statistical tests.
-- Appropriate recognition and treatment of missing data issues.
-- Appropriate and effective use of tables, charts, and graphs: i.e., they complement the analysis in your memo; convey information clearly; are appropriately labeled and formatted.
-- Ability to address both audiences (1) and (2) described above.
-- Overall style, organization, and clarity of the argument / analysis.

**IF YOU HAVE QUESTIONS ABOUT THE DATA STRUCTURE, SPECIFIC VARIABLES, WHAT YOU ARE BEING ASKED TO DO, ETC., PLEASE ASK!**